PATENT

Attorney Docket No. 16528X-86
Affymetrix Reference No. 1079

# *SEQUENCING BY HYBRIDIZATION ON HIGH DENSITY PROBE ARRAYS: ENZYMATIC DISCRIMINATION ENHANCEMENT*

## BACKGROUND OF THE INVENTION

The relationship between structure and function of macromolecules is of fundamental importance in the understanding of biological systems. Such relationships are important to understanding, for example, the functions of enzymes, structural proteins, and signalling proteins, the ways in which cells communicate with one another, the mechanisms of cellular control and metabolic feedback, *etc.*

Genetic information is critical in continuation of life processes. Life is substantially informationally based, and genetic content controls the growth and reproduction of the organism and its complements. Proteins, which are critical features of all living systems, are encoded by the genetic materials of the cell. More particularly, the properties of enzymes, functional proteins, and structural proteins are determined by the sequence of amino acids from which they are made. As such, it has become very important to determine the genetic sequences of nucleotides which encode the enzymes, structural proteins and other effectors of biological functions. In addition to the segments of nucleotides which encode polypeptides, there are many nucleotide sequences which are involved in the control and regulation of gene expression.

The human genome project is an example of a project that is directed toward determining the complete sequence of the genome of the human organism. Although such a sequence would not necessarily correspond to the sequence of any specific individual, it will provide significant information as to the general organization and specific sequences contained within genomic segments from particular individuals. It will also provide mapping information useful for further detailed studies. The need for highly rapid, accurate, and inexpensive sequencing technology is nowhere more apparent than in a demanding sequencing project such as this. To complete the sequencing of a human genome will require the determination of approximately 3 x $10^9$, or 3 billion, base pairs.

The procedures typically used today for sequencing include the methods described in Sanger, *et al.*, *Proc. Natl. Acad. Sci. USA* **74**:5463-5467 (1977), and Maxam, *et al.*, *Methods in Enzymology* **65**:499-559 (1980). The Sanger method utilizes enzymatic elongation with chain terminating dideoxy nucleotides. The Maxam and Gilbert method uses chemical reactions exhibiting specificity of reactants to generate nucleotide specific cleavages. Both methods, however, require a practitioner to perform a large number of complex, manual manipulations. For example, such methods usually require the isolation of homogeneous DNA fragments, elaborate and tedious preparation of samples, preparation of a separating gel, application of samples to the gel, electrophoresing the samples on the gel, working up the finished gel, and analysis of the results of the procedure.

Alternative techniques have been proposed for sequencing a nucleic acid. PCT patent Publication No. 92/10588, incorporated herein by reference for all purposes, describes one improved technique in which the sequence of a labeled, target nucleic acid is determined by hybridization to an array of nucleic acid probes on a substrate. Each probe is located at a positionally distinguishable location on the substrate. When the labeled target is exposed to the substrate, it binds at locations that contain complementary nucleotide sequences. Through knowledge of the sequence of the probes at the binding locations, one can determine the nucleotide sequence of the target nucleic acid. The technique is particularly efficient when very large arrays of nucleic acid probes are utilized. Such arrays can be formed according to the techniques described in U.S. Patent No. 5,143,854 issued to Pirrung, *et al.* *See also*, U.S. application Serial No. 07/805,727, both of which are incorporated herein by reference for all purposes.

When the nucleic acid probes are of a length shorter than the target, one can employ a reconstruction technique to determine the sequence of the larger target based on affinity data from the shorter probes. *See*, U.S. Patent No. 5,202,231 issued to Drmanac, *et al.*, and PCT patent Publication No. 89/10977 issued to Southern. One technique for overcoming this difficulty has been termed sequencing by hybridization or SBH. Assume, for example, that a 12-mer target DNA, *i.e.*, 5'-AGCCTAGCTGAA, is mixed with an array of all octanucleotide probes. If the target binds only to those probes having an exactly complementary nucleotide sequence, only five of the 65,536 octamer probes (*i.e.*, 3'-TCGGATCG, CGGATCGA, GGATCGAC, GATCGACT, and

ATCGACTT) will hybridize to the target. Alignment of the overlapping sequences from the hybridizing probes reconstructs the complement of the original 12-mer target:

TCGGATCG
CGGATCGA
GGATCGAC
GATCGACT
ATCGACTT
TCGGATCGACTT (Seq. ID NO.:1)

Although such techniques have been quite useful, it would be helpful to have additional methods which can effectively discriminate between fully complementary hybrids and those that differ by one or more base pairs. Quite surprisingly, the present invention provides such methods.

## SUMMARY OF THE INVENTION

The present invention provides improved methods for discriminating between fully complementary hybrids and those that differ by one or more base pairs. In one embodiment, the present invention provides methods of using nuclease treatment to improve the quality of hybridization signals on high density oligonucleotide arrays. More particularly, in this method, an array of oligonucleotides is combined with a labelled target nucleic acid to form target-oligonucleotide hybrid complexes. Thereafter, the target-oligonucleotide hybrid complexes are treated with a nuclease and, in turn, the array of target-oligonucleotide complexes are washed to remove non-perfectly complementary target-oligonucleotide hybrid complexes. Following nuclease treatment, the target:oligonucelotide hybrid complexes which are perfectly complementary are identified. From the location of the labelled targets, the oligonucleotide probes which hybridized with the targets can be identified and, in turn, the sequence of the target nucleic acid can be determined.

In another embodiment, the present invention provides methods wherein ligiation reactions are used to discriminate between fully complementary hybrids and those that differ by one or more base pairs. In this method, an array of oligonucelotides is generated on a substrate (in the 3' to 5' direction) using any one of the methods described above. Each of the oligonucleotides in the array is shorter in length than the target nucleic acid so that when hybridized to the target nucleic acid, the target nucleic

acid generally has a 3' overhang. In this embodiment, the target nucleic acid is not necessarily labelled. After the array of oligonucelotides has been combined with the target nucleic acid to form target-oligonucleotide hydrid complexes, the target-oligonucleotide hybrid complexes are contacted with a ligase and a labelled, ligatable probe or, alternatively, with a pool of labelled, ligatable probes. The ligation reaction of the labelled, ligatable probes to the 5' end of the oligonucleotide probes on the substrate will occur, in the presence of the ligase, only when the target:oligonucleotide hybrid has formed with correct base-pairing near the 5' end of the oligonucleotide probe and where there is a suitable 3' overhang of the target nucleic acid to serve as a template for hybridization and ligation. After the ligation reaction, the substrate is washed (multiple times if necessary) with water at a temperature of about 40°C to 50°C to remove the target nucleic acid and the labelled, unligated probes. Thereafter, a quantitative fluorescence image of the hybridization pattern is obtained by scanning the substrate with, for example, a confocal microscope, and labelled oligonucleotide probes, *i.e.*, the oligonucleotide probes which are perfectly complementary to the target nucleic acid, are identified. Using this information, the sequence of the target nucleic acid can be determined.

A further understanding of the nature and advantages of the inventions herein may be realized by reference to the remaining portions of the specification and the attached drawings.

## BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 illustrates discrimination of non-perfectly complementary target:oligonucleotide hybrids using RNase A.

FIG. 2 illustrates discrimination of non-perfectly complementary target:oligonucleotide hybrids using a ligation reaction.

FIG. 3 illustrates the light directed synthesis of an array of oligonucleotides on a substrate.

FIG. 4. illustrates a hybridization procedure which can be used prior to nuclease treatment.

FIG. 5 illustrates probe tiling strategy used to generate the probes.

FIG. 6 illustrates the results obtained from hybridization to the substrate without RNase treatment.

FIG. 6 illustrates the results obtained from hybridization to the substrate with RNase treatment.

## DETAILED DESCRIPTION OF THE INVENTION AND PREFERRED EMBODIMENT

### TABLE OF CONTENTS

## *I. Glossary*

The following terms are intended to have the following general meanings as they are used herein:

1. Substrate: A material having a rigid or semi-rigid surface. In many embodiments, at least one surface of the substrate will be substantially flat, although in some embodiments it may be desirable to physically separate synthesis regions for different polymers with, for example, wells, raised regions, etched trenches, or the like. In some embodiments, the substrate itself contains wells, trenches, flow through regions, *etc.* which form all or part of the synthesis regions. According to other embodiments, small beads may be provided on the surface, and compounds synthesized thereon may be released upon completion of the synthesis.

2. Predefined Region: A predefined region is a localized area on a substrate which is, was, or is intended to be used for formation of a selected polymer and is otherwise referred to herein in the alternative as "reaction" region, a "selected" region, or simply a "region." The predefined region may have any convenient shape, *e.g.*, circular, rectangular, elliptical, wedge-shaped, *etc.* In some embodiments, a predefined region and, therefore, the area upon which each distinct polymer sequence is synthesized is smaller than about 1 $cm^2$, more preferably less than 1 $mm^2$, and still more preferably less than 0.5 $mm^2$. In most preferred embodiments, the regions have an area less than about 10,000 $\mu m^2$ or, more preferably, less than 100 $\mu m^2$. Within these regions, the polymer synthesized therein is preferably synthesized in a substantially pure form.

3. Substantially Pure: A polymer or other compound is considered to be "substantially pure" when it exhibits characteristics that distinguish it from the polymers or compounds in other regions. For example, purity can be measured in terms of the activity or concentration of the compound of interest. Preferably the compound in a region is sufficiently pure such that it is the predominant species in the region. According to certain aspects of the invention, the compound is 5% pure, more preferably more than 10% pure, and most preferably

more than 20% pure. According to more preferred aspects of the invention, the compound is greater than 80% pure, preferably more than 90% pure, and more preferably more than 95% pure, where purity for this purpose refers to the ratio of the number of compound molecules formed in a region having a desired structure to the total number of non-solvent molecules in the region.

4. Monomer: In general, a monomer is a member of a set of small molecules which are or can be joined together to form a polymer. The particular ordering of monomers within a polymer is referred to herein as the "sequence" of the polymer. As used herein, monomer refers to any member of a basis set used for synthesis of a polymer. For example, dimers of the 20 naturally occurring L-amino acids form a basis set of 400 monomers for synthesis of polypeptides. Different basis sets of monomers may be used at successive steps in the synthesis of a polymer. Furthermore, each of the sets may include protected members which are modified after synthesis. The invention described herein can readily be applied in the preparation and screening of diverse types of polymers. Such polymers include, for example, both linear and cyclic polymers of nucleic acids, polysaccharides, phospholipids, and peptides having either $\alpha$-, $\beta$-, or $\omega$-amino acids, heteropolymers in which a known drug is covalently bound to any of the above, polyacetates, polyamides, polyarylene sulfides, polycarbamates, polycarbonates, polyesters, polyethyleneimines, polyimides, polynucleotides, polyphosphonates, polysiloxanes, polysulfones, polysulfoxides, polyureas, polyurethanes, or other polymers which will be apparent upon review of this disclosure.

5. Protective Group: A material which is bound to a monomer or other compound or group and which may be selectively removed therefrom to expose an active site such as, in the example of an amino acid, an amine group. A protective group will typically be used to block one reactive site of a bifunctional monomer from reacting during an addition reaction such as formation of a peptide from amino acids. A protective group can also cover certain regions of a substrate surface to impart certain properties such as non-wettability and to define region perimeters or other features.

6. Complementary or substantially complementary: Refers to the hybridization or base pairing between nucleotides or nucleic acids, such as, for instance, between the two strands of a double stranded DNA molecule, or between an oligonucleotide primer and a primer binding site on a single stranded nucleic acid to be sequenced or amplified. Complementary nucleotides are, generally, A and T (or A and U), or C and G. Two single stranded RNA or DNA molecules are said to be substantially complementary when the nucleotides of one strand, optimally aligned and compared and with appropriate nucleotide insertions or deletions, pair with at least about 80% of the nucleotides of the other strand, usually at least about 90% to 95%, and more preferably from about 98 to 100%. Alternatively, substantial complementarity exists when an RNA or DNA strand will hybridize under selective hybridization conditions to its complement. Typically, selective hybridization will occur when there is at least about 65% complementarity over a stretch of at least 14 to 25 nucleotides, preferably at least about 75%, more preferably at least about 90% complementarity. *See*, M. Kanehisa *Nucleic Acids Res.* **12**:203 (1984), incorporated herein by reference.

7. Stringent hybridization conditions: Such conditions will typically include salt concentrations of less than about 1 M, more usually less than about 500 mM, and preferably less than about 200 mM. Hybridization temperatures can be as low as 5°C, but are typically greater than 22°C, more typically greater than about 30°C, and preferably in excess of about 37°C. Longer fragments may require higher hybridization temperatures for specific hybridization. As other factors may dramatically affect the stringency of hybridization, including base composition, length of the complementary strands, presence of organic solvents and extent of base mismatching, the combination of parameters is more important than the absolute measure of any one alone.

8. Oligonucleotides: An oligonucleotide is a single-stranded DNA or RNA molecule, typically prepared by synthetic means. Alternatively, naturally occurring oligonucleotides, or fragments thereof, may be isolated from their natural sources or purchased from commercial sources. Those oligonucleotides employed in the present invention will be 4 to 100 nucleotides in length,

preferably from 6 to 30 nucleotides, although oligonucleotides of different length may be appropriate. Suitable oligonucleotides may be prepared by the phosphoramidite method described by Beaucage and Carruthers, *Tetrahedron Lett.*, **22**:1859-1862 (1981), or by the triester method according to Matteucci, *et al.*, *J. Am. Chem. Soc.*, **103**:3185 (1981), both incorporated herein by reference, or by other chemical methods using either a commercial automated oligonucleotide synthesizer or VLSIPS™ technology (discussed in detail below). When oligonucleotides are referred to as "double-stranded," it is understood by those of skill in the art that a pair of oligonucleotides exist in a hydrogen-bonded, helical array typically associated with, for example, DNA. In addition to the 100% complementary form of double-stranded oligonucleotides, the term "double-stranded" as used herein is also meant to refer to those forms which include such structural features as bulges and loops, described more fully in such biochemistry texts as Stryer, *Biochemistry*, Third Ed., (1988), previously incorporated herein by reference for all purposes.

9. Probe: A molecule of known composition or monomer sequence, typically formed on a solid surface, which is or may be exposed to a target molecule and examined to determine if the probe has hybridized to the target. Also referred to herein as an "oligonucleotide" or an "oligonucleotide probe."

10. Target: A molecule, typically of unknown composition or monomer sequence, for which it is desired to study the composition or monomer sequence. A target may be a part of a larger molecule, such as a few bases in a longer nucleic acid.

11. A, T, C, G, U: A, T, C, G, and U are abbreviations for the nucleotides adenine, thymine, cytosine, guanine, and uridine, respectively.

12. Array or Library: A collection of oligonucleotide probes of predefined nucleotide sequence, often formed in one or more substrates, which are used in hybridization studies of target nucleic acids.

## *II. General Overview*

The present invention provides improved methods for obtaining sequence information about nucleic acids (*i.e.*, oligonucleotides). More particularly, the present invention provides improved methods for discriminating between fully complementary hybrids and those that differ by one or more base pairs. The methods of the present invention rely, in part, on the ability to synthesize or attach specific oligonucleotides at known locations on a substrate, typically a single substrate. Such oligonucleotides are capable of interacting with specific target nucleic acid while attached to the substrate. By appropriate labeling of these targets, the sites of the interactions between the target and the specific oligocnucleotide can be derived. Moreover, because the oligonucleotides are positionally defined, the target sequence can be reconstructed from the sites of the interactions.

It has now been determined that reconstruction of the target sequence can be improved by using various enzymes that catalyze oligonucleotide cleavage and ligation reactions. More particularly, it has been determined that discrimination between fully complementary hybrids and those that differ by one or more base pairs can be greatly enhanced by using various enzymes that catalyze oligonucleotide cleavage and ligation reactions.

RNase A treatment, for example, can be used to improve the quality of RNA hybridization signals on high density oligonucleotide arrays. After the array of oligonucleotides has been combined with a target nucleic acid (RNA) to form target-oligonucleotide hybrid complexes, the target-oligonucleotide hybrid complexes are treated with RNase A to remove non-perfectly complementary target-oligonucleotide hybrid complexes. RNase A recognizes and cuts single-stranded RNA, including RNA in RNA:DNA hybrids that is not in a perfect double-stranded structure. As illustrated in FIG. 1, RNA bulges, loops, and even single base mismatches can be recognized and cleaved by RNase A. Similarly, treatment with other nucleases (*e.g.*, S1 nuclease and Mung Bean nuclease) can be used to improve the DNA hybridization signals on high density oligonucleotide arrays. As such, nuclease treatment can be used to improve the quality of hybridization signals on high density oligonucleotide arrays and, in turn, to more accurately determine the sequence of the target nucleic acid.

Moreover, ligation reactions can be used to discriminate between fully complementary hybrids and those that differ by one or more base pairs. T4 DNA ligase,

for example, can be used to identify DNA:DNA hybrids that are perfectly complementary near the 5' end of the immobilized oligonucleotide probes. The ligation reaction of labelled, short oligonucleotides to the 5' end of oligonucleotide probes on a substrate will occur, in the presence of the enzyme Ligase, only when a target:oligonucleotide hybrid has formed with correct base-pairing near the 5' end of the oligonucleotide probe and where there is a suitable 3' overhang of the target to serve as a template for hybridization and ligation. As such, after the array of oligonucleotides has been combined with a target nucleic acid to form target-oligonucleotide hybrid complexes, the target-oligonucleotide hybrid complexes can be contacted with a ligase and a labelled, ligatable oligonucleotide probe. After the ligation reaction, the substrate is washed to remove the target nucleic acid and labelled, unligatable oligonucleotide probes. The oligonucleotide probes containing the label indicate sequences which are perfectly complementary to target nucleic acid sequence. As such, as illustrated in FIG. 2, ligation reactions can be used to improve discrimination of base-pair mismatches near the 5' end of the probe, mismatches that are often poorly discriminated following hybridization alone.

## *III. Methods For Generating An Array Of Oligonucleotides On A Substrate*

### *A. The Substrate*

In the methods of the present invention, an array of diverse oligonucleotides at known locations on a single substrate surface is employed. Essentially, any conceivable substrate can be employed in the invention. The substrate can be organic, inorganic, biological, nonbiological, or a combination of any of these, existing as beads, particles, strands, precipitates, gels, sheets, tubing, spheres, containers, capillaries, pads, slices, films, plates, slides, *etc.* The substrate can have any convenient shape, such a disc, square, sphere, circle, *etc.* The substrate is preferably flat, but may take on a variety of alternative surface configurations. For example, the substrate may contain raised or depressed regions on which the synthesis takes place. The substrate and its surface preferably form a rigid support on which to carry out the reaction described herein. The substrate and its surface may also chosen to provide appropriate light-absorbing characteristics. The substrate may be any of a wide variety of materials including, for example, polymers, plastics, pyrex, quartz, resins, silicon, silica or silica-based materials, carbon, metals, inorganic glasses, inorganic crystals,

membranes, *etc.* More particularly, the substrate may, for instance, be a polymerized Langmuir Blodgett film, functionalized glass, Si, Ge GaAs, GaP, $SiO_2$, $SiN_4$, modified silicon, or any one of a wide variety of gels or polymers such as (poly)-tetrafluorotheylene, (poly)vinylidenedifluoride, polystyrene, polycarbonate, or combinations thereof. Other substrate materials will be readily apparent to those of skill in the art upon review of this disclosure. In a preferred embodiment the substrate is flat glass or single-crystal silicon with surface relief features of less than 10.

In some embodiments, a predefined region on the substrate and, therefore, the area upon which each distinct material is synthesized will have a surface area of between about 1 $cm^2$ and $10^{-10}cm^2$. In some embodiments, the regions have areas of less than about $10^{-1}cm^2$, $10^{-2}cm^2$, $10^{-3}cm^2$, $10^{-4}cm^2$, $10^{-5}cm^2$, $10^{-6}cm^2$, $10^{-7}cm^2$, $10^{-8}cm^2$, or $10^{-10}cm^2$. In a preferred embodiment, the regions are between about 10X10 $\mu$m and 500x100$\mu$m.

Moreover, in some embodiments, a single substrate supports more than about 10 different monomer sequences and preferably more than about 100 different monomer sequences, although in some embodiments more than about $10^3$, $10^4$, $10^5$, $10^6$, $10^7$, or $10^8$ different sequences are provided on a substrate. Of course, within a region of the substrate in which a monomer sequence is synthesized, it is preferred that the monomer sequence be substantially pure. In some embodiments, regions of the substrate contain polymer sequences which are at least about 1%, 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 60%, 70%, 80% 90%, 95%, 96%, 97%, 98%, or 99% pure.

As previously explained, the substrate is preferably flat, but may take on a variety of alternative surface configurations. Regardless of the configuration of the substrate surface, it is imperative that the reactants used to generate an array of oligonucleotides in the individual reaction regions be prevented from moving to adjacent reaction regions. Most simply, this is ensured by chemically attaching the oligonucleotides to the substrate. Moreover, this can be ensured by providing an appropriate barrier between the various reaction regions on the substrate. A mechanical device or physical structure can be used to define the various regions on the substrate. For example, a wall or other physical barrier can be used to prevent the reactants in the individual reaction regions from moving to adjacent reaction regions. Alternatively, a

dimple or other recess can be used to prevent the reactant components in the individual reaction regions from moving to adjacent reaction regions.

### *B. Generating An Array Using Light-Directed Methods*

An array of diverse oligonucleotides diverse oligonucleotides at known locations on a single substrate surfaces can be formed using a variety of techniques known to those skilled in the art of polymer synthesis on solid supports. For example, "light directed" methods (which are one technique in a family of methods known as VLSIPS™ methods) are described in U.S. Patent No. 5,143,854, previously incorporated by reference. The light directed methods discussed in the '854 patent involve activating predefined regions of a substrate or solid support and then contacting the substrate with a preselected monomer solution. The predefined regions can be activated with a light source shown through a mask (much in the manner of photolithography techniques used in integrated circuit fabrication). Other regions of the substrate remain inactive because they are blocked by the mask from illumination and remain chemically protected. Thus, a light pattern defines which regions of the substrate react with a given monomer. By repeatedly activating different sets of predefined regions and contacting different monomer solutions with the substrate, a diverse array of polymers is produced on the substrate. Of course, other steps such as washing unreacted monomer solution from the substrate can be used as necessary. Other techniques include mechanical techniques such as those described in PCT No. 92/10183, USSN 07/796,243, also incorporated herein by reference for all purposes. Still further techniques include bead based techniques such as those described in PCT US/93/04145, also incorporated herein by reference, and pin based methods such as those described in U.S. Pat. No. 5,288,514, also incorporated herein by reference.

The VLSIPS™ methods are preferred for generating an array of oligonucleotides on a single substrate. As illustrated in FIG. 1, the surface of a solid support, optionally modified with spacers having photolabile protecting groups such as NVOC and MeNPOC, is illuminated through a photolithographic mask, yielding reactive groups (typically hydroxyl groups) in the illuminated regions. A 3'-O-phosphoramidite activated deoxynucleoside (protected at the 5'-hydroxyl with a photolabile protecting group) is then presented to the surface and chemical coupling occurs at sites that were exposed to light. Following capping, and oxidation, the substrate is rinsed and the

surface illuminated through a second mask, to expose additional hydroxyl groups for coupling. A second 5'-protected, 3'-O-phosphoramidite activated deoxynucleoside is presented to the surface. The selective photodeprotection and coupling cycles are repeated until the desired set of oligonucleotides is produced.

### *B. Generating An Array Of Oligonucleotides Using Flow Channel Or Spotting Methods*

In addition to the foregoing, additional methods which can be used to generate an array of oligonucleotides on a single substrate are described in co-pending Applications Ser. No. 07/980,523, filed November 20, 1992, and 07/796,243, filed November 22, 1991, incorporated herein by reference for all purposes. In the methods disclosed in these applications, reagents are delivered to the substrate by either (1) flowing within a channel defined on predefined regions or (2) "spotting" on predefined regions. However, other approaches, as well as combinations of spotting and flowing, may be employed. In each instance, certain activated regions of the substrate are mechanically separated from other regions when the monomer solutions are delivered to the various reaction sites.

A typical "flow channel" method applied to the compounds and libraries of the present invention can generally be described as follows. Diverse polymer sequences are synthesized at selected regions of a substrate or solid support by forming flow channels on a surface of the substrate through which appropriate reagents flow or in which appropriate reagents are placed. For example, assume a monomer "A" is to be bound to the substrate in a first group of selected regions. If necessary, all or part of the surface of the substrate in all or a part of the selected regions is activated for binding by, for example, flowing appropriate reagents through all or some of the channels, or by washing the entire substrate with appropriate reagents. After placement of a channel block on the surface of the substrate, a reagent having the monomer A flows through or is placed in all or some of the channel(s). The channels provide fluid contact to the first selected regions, thereby binding the monomer A on the substrate directly or indirectly (via a spacer) in the first selected regions.

Thereafter, a monomer B is coupled to second selected regions, some of which may be included among the first selected regions. The second selected regions will be in fluid contact with a second flow channel(s) through translation, rotation, or

replacement of the channel block on the surface of the substrate; through opening or closing a selected valve; or through deposition of a layer of chemical or photoresist. If necessary, a step is performed for activating at least the second regions. Thereafter, the monomer B is flowed through or placed in the second flow channel(s), binding monomer B at the second selected locations. In this particular example, the resulting sequences bound to the substrate at this stage of processing will be, for example, A, B, and AB. The process is repeated to form a vast array of sequences of desired length at known locations on the substrate.

After the substrate is activated, monomer A can be flowed through some of the channels, monomer B can be flowed through other channels, a monomer C can be flowed through still other channels, *etc.* In this manner, many or all of the reaction regions are reacted with a monomer before the channel block must be moved or the substrate must be washed and/or reactivated. By making use of many or all of the available reaction regions simultaneously, the number of washing and activation steps can be minimized.

One of skill in the art will recognize that there are alternative methods of forming channels or otherwise protecting a portion of the surface of the substrate. For example, according to some embodiments, a protective coating such as a hydrophilic or hydrophobic coating (depending upon the nature of the solvent) is utilized over portions of the substrate to be protected, sometimes in combination with materials that facilitate wetting by the reactant solution in other regions. In this manner, the flowing solutions are further prevented from passing outside of their designated flow paths.

The "spotting" methods of preparing compounds and libraries of the present invention can be implemented in much the same manner as the flow channel methods. For example, a monomer A can be delivered to and coupled with a first group of reaction regions which have been appropriately activated. Thereafter, a monomer B can be delivered to and reacted with a second group of activated reaction regions. Unlike the flow channel embodiments described above, reactants are delivered by directly depositing (rather than flowing) relatively small quantities of them in selected regions. In some steps, of course, the entire substrate surface can be sprayed or otherwise coated with a solution. In preferred embodiments, a dispenser moves from region to region, depositing only as much monomer as necessary at each stop. Typical dispensers include a micropipette to deliver the monomer solution to the substrate and a robotic system to

control the position of the micropipette with respect to the substrate. In other embodiments, the dispenser includes a series of tubes, a manifold, an array of pipettes, or the like so that various reagents can be delivered to the reaction regions simultaneously.

### *C. Generating An Array Of Oligonucleotides Using Pin-Based Methods*

Another method which is useful for the preparation of an array of diverse oligonucleotides on a single substrate involves "pin based synthesis." This method is described in detail in U.S. Patent No. 5,288,514, previously incorporated herein by reference. The method utilizes a substrate having a plurality of pins or other extensions. The pins are each inserted simultaneously into individual reagent containers in a tray. In a common embodiment, an array of 96 pins/containers is utilized.

Each tray is filled with a particular reagent for coupling in a particular chemical reaction on an individual pin. Accordingly, the trays will often contain different reagents. Since the chemistry used is such that relatively similar reaction conditions may be utilized to perform each of the reactions, multiple chemical coupling steps can be conducted simultaneously. In the first step of the process, a substrate on which the chemical coupling steps are conducted is provided. The substrate is optionally provided with a spacer having active sites. In the particular case of oligonucleotides, for example, the spacer may be selected from a wide variety of molecules which can be used in organic environments associated with synthesis as well as in aqueous environments associated with binding studies. Examples of suitable spacers are polyethyleneglycols, dicarboxylic acids, polyamines and alkylenes, substituted with, for example, methoxy and ethoxy groups. Additionally, the spacers will have an active site on the distal end. The active sites are optionally protected initially by protecting groups. Among a wide variety of protecting groups which are useful are FMOC, BOC, t-butyl esters, t-butyl ethers, and the like. Various exemplary protecting groups are described in, for example, Atherton *et al.*, *Solid Phase Peptide Synthesis,* IRL Press (1989), incorporated herein by reference. In some embodiments, the spacer may provide for a cleavable function by way of, for example, exposure to acid or base.

### *D. Generating An Array Of Oligonucleotides Using Bead Based Methods*

In addition to the foregoing methods, another method which is useful for synthesis of an array of oligonucleotids involves "bead based synthesis." A general approach for bead based synthesis is described in copending Application Ser. Nos. 07/762,522 (filed September 18, 1991); 07/946,239 (filed September 16, 1992); 08/146,886 (filed November 2, 1993); 07/876,792 (filed April 29, 1992) and PCT/US93/04145 (filed April 28, 1993), the disclosures of which are incorporated herein by reference.

For the synthesis of molecules such as oligonucleotides on beads, a large plurality of beads are suspended in a suitable carrier (such as water) in a container. The beads are provided with optional spacer molecules having an active site. The active site is protected by an optional protecting group.

In a first step of the synthesis, the beads are divided for coupling into a plurality of containers. For the purposes of this brief description, the number of containers will be limited to three, and the monomers denoted as A, B, C, D, E, and F. The protecting groups are then removed and a first portion of the molecule to be synthesized is added to each of the three containers (*i.e.*, A is added to container 1, B is added to container 2 and C is added to container 3).

Thereafter, the various beads are appropriately washed of excess reagents, and remixed in one container. Again, it will be recognized that by virtue of the large number of beads utilized at the outset, there will similarly be a large number of beads randomly dispersed in the container, each having a particular first portion of the monomer to be synthesized on a surface thereof.

Thereafter, the various beads are again divided for coupling in another group of three containers. The beads in the first container are deprotected and exposed to a second monomer (D), while the beads in the second and third containers are coupled to molecule portions E and F, respectively. Accordingly, molecules AD, BD, and CD will be present in the first container, while AE, BE, and CE will be present in the second container, and molecules AF, BF, and CF will be present in the third container. Each bead, however, will have only a single type of molecule on its surface. Thus, all of the possible molecules formed from the first portions A, B, C, and the second portions D, E, and F have been formed.

The beads are then recombined into one container and additional steps are conducted to complete the synthesis of the polymer molecules. In a preferred embodiment, the beads are tagged with an identifying tag which is unique to the particular oligonucleotide which is present on each bead. A complete description of identifier tags for use in synthetic libraries is provided in co-pending Application Ser. No. 08/146,886 (filed November 2, 1993), previously incorporated by reference for all purposes.

## *IV. Sequencing By Hybridization*

The principle of the hybridization sequencing procedure used in the methods of the present invention is based, in part, upon the ability to determine overlaps of short segments. The VLSIPS™ technology provides the ability to generate an array of oligonucleotides which will saturate the possible short subsequence recognition possibilities. This principle is most easily illustrated by using a binary sequence, such as a sequence of zeros and ones. Once having illustrated the application to a binary alphabet, the principle can easily be understood to encompass three letter, four letter, five letter, or even 20 letter alphabets. A theoretical treatment of analysis of subsequence information to reconstruction of a target sequence is provided, *e.g.*, in Lysov, Yu., *et al.*, *Doklady Akademi. Nauk. SSR* 303:1508-1511 (1988); Khropko K., *et al. FEBS Letters* 256:118-122 (1989); Pevzner, P. (1929) *J. of Biomolecular Structure and Dynamics* 7:63-69; and Drmanac, R., *et al.*, *Genomics* 4:114-128 (1989); each of which is hereby incorporated herein by reference.

The oligonucleotides used for recognizing the subsequences will usually be specific for a particular nucleic acid subsequence anywhere within a given target. Using these hybridization sequencing techniques, in combination with the nuclease and ligation methods disclosed herein, improved discrimination between high fidelity matching and very low levels of mismatching can be achieved.

### *A. Simple n-mer Structure: Theory*

#### 1. Example of Using A Two Letter Alphabet

A simple example is presented below of how a sequence of ten digits, comprising zeros and ones, can be sequenced using short oligonucleotides of five digits. Consider, for example, the following ten digit sequence:

1010011100.

A VLSIPS™ substrate can be generated, using the methods described above, which would have an array of oligonucleotides attached in a defined matrix pattern which specifically recognize each of the possible five digit sequences of ones and zeros. The number of possible five digit subsequences is $2^5 = 32$. The number of possible ten digit long sequences is $2^{10} = 1{,}024$. The number of contiguous five digit subsequences within a ten digit long sequence is six, *i.e.*, positioned at digits 1-5, 2-6, 3-7, 4-8, 5-9, and 6-10. It will be noted that the specific order of the digits in the sequence is important and that the order is directional, *e.g.*, running left to right versus right to left. The first five digit sequence contained in the target sequence is 10100; the second is 01001; the third is 10011; the fourth is 00111; the fifth is 01110; and the sixth is 11100.

As such, a VLSIPS™ substrate would be generated to have a matrix pattern of positionally attached oligonucleotides which recognize each of the different 5-mer subsequences. The oligonucleotides which recognize each of the 6 contiguous 5-mer subsequences will bind to the target, and a label is used to identify the positional determination of where a sequence specific interaction has occurred. By correlation of the position in the matrix pattern, the corresponding bound subsequences can be determined.

In the above-mentioned sequence, six different contiguous 5-mer sequences would be determined to be present. They would be:

10100
01001
10011
00111
01110
11100

From an analysis of the foregoing subsequences, it is seen that any sequence which contains the first five digit subsequence, *i.e.*, 10100, already narrows the number of possible sequences which contain it from 1024 possible sequences to less than

about 192 possible sequences. This 192 is derived from the observation that with the subsequence 10100 at the far, left of the sequence, in positions 1-5, there are only 32 possible sequences. Likewise, for that particular subsequence in positions 2-6, 3-7, 4-8, 5-9, and 6-10. Thus, in summing up all of the sequences that can contain 10100, there are 32 for each position and, thus, for 6 positions, a total of about 192 possible sequences is obtained. It should be noted that some of the 10 digit sequences will have been counted twice. Thus, by virtue of containing the 10100 subsequence, the number of possible 10-mer sequences has been decreased form 1024 sequences to less than about 192 sequences.

In this example, not only is it known that the sequence contains 10100, but it is also know that it contains the second five character sequence, *i.e.*, 01001. By virtue of knowing that the sequence contains 10100, one can look specifically to determine whether the sequence contains a subsequence of five characters which contains the four leftmost digits plus a next digit to the left. For example, one can specifically look for a sequence of X1010, but, in doing so, one would find that there is no such sequence. Thus, it is determined that the 10100 must be at the left end of the 10-mer. In addition, one would want to look to see whether the sequence contains the rightmost four digits plus a next digit to the right, *e.g.*, 0100X. In doing so, it is found that the sequence contains the sequence 01001, and that X is a 1. Thus, it is known that the target sequence has an overlap of 0100 and has the left terminal sequence 101001.

Applying the some procedure to the second 5-mer, it is also known that the sequence must include a subsequence of five digits having the sequence 1001Y, where Y must be either 0 or 1. Looking through the fragments obtained, it is seen that there is a 10011 sequence within the target, thus Y is also 1. Thus, it is known that the sequence contains has a subsequence in which the first seven digits are: 1010011.

Moving to the next 5-mer, it is known that there must be a sequence of 0011Z, where Z must be either 0 or 1. Looking at the fragments obtained, it is seen that the target sequence contains a 00111 subsequence and Z is 1. Thus, it is known that the sequence must start with 10100111. The next 5-mer must be of the sequence 0111W, where W must be 0 or 1. Again, looking at the fragments obtained, it is seen that the target sequence contains a 01110 subsequence, and W is a 0. Thus, the subsequence identified up to this point is 101001110.

Finally, it is known that the last 5-mer must be either 11100 or 11101. Looking at the remaining fragments, it is seen that the last of the 5-mer subsequences 11100. Thus, the target sequence has been determined to have the following sequence: 1010011100.

It will be recognized from the example above with the sequences provided therein, however, that the sequence analysis can start with any known positive oligonucleotide subsequence. Sequence determination can be performed by moving linearly along the sequence checking the known sequence with a limited number of next positions. Given this possibility, in addition to scanning all possible oligonucleotide positions, the sequence may be determined by specifically looking at only where the next possible positions would be. This may increase the complexity of the scanning, but may provide a longer time span dedicated towards scanning and detecting specific positions of interest relative to other sequence possibilities. Thus, the scanning apparatus could be set up to work its way along a sequence from a given contained oligonucleotide to only look at those positions on the substrate which are expected to have a positive signal.

It is seen that given a sequence, it can be deconstructed into n-mers to produce a set of internal contiguous subsequences. From any given target sequence, one is able to determine what fragments would result. The hybridization sequence method depends, in part, upon being able to work in the reverse, from a set of fragments of known sequences to the full sequence. In simple cases, one is able to start at a single position and work in either or both directions towards the ends of the sequence as illustrated in the example.

The number of possible sequences of a given length increases very quickly with the length of that sequence. Thus, a 10-mer of zeros and ones has 1024 possibilities, a 12-mer has 4096. A 20-mer has over a million possibilities, and a 30-mer has over a billion. However, a given 30-mer has, at most, 26 different internal 5-mer sequences. Thus, a 30 character target sequence having over a million possible sequences can be substantially defined by only 26 different 5-mers. It will be recognized that the oligonucleotides will preferably, but need not necessarily, be of identical length, and that the oligonucleotide sequences need not necessarily be contiguous in that the overlapping subsequences need not differ by only a single subunit. Moreover, each position of the matrix patten need not be homogenous, but may actually contain a plurality of probes of known sequence. In addition, although all of the possible

subsequences specifications would be preferred, a less than full set of sequences specifications can be used. In particular, although a substantial fraction will preferably be at least about 70%, it may be less than that. Although higher percentages are especially preferred, at least about 20%, more preferably at least about 30%, can be used.

## *V. Enzymatic Discrimination Enhancement*

Unfortunately using the foregoing techniques, it is frequently difficult to discriminate between fully complementary hybrids and those that differ by one or more base pairs. However, it has now been determined that sequencing by hybridization can be improved by using various enzymes that catalyze oligonucleotide cleavage and ligation reactions. More particularly, discrimination between fully complementary hybrids and those that differ by one or more base pairs can be greatly enhanced by using various enzymes that catalyze oligonucleotide cleavage and ligation reactions.

### *A. Enhanced Discrimination Using Nuclease Treatment*

Nuclease treatment can be used to improve the quality of hybridization signals on high density oligonucleotide arrays. More particularly, after the array of oligonucleotides has been combined with a labelled target nucleic acid to form target-oligonucleotide hybrid complexes, the target-oligonucleotide hybrid complexes are treated with a nuclease and, in turn, they are washed to remove non-perfectly complementary target-oligonucleotide hybrid complexes. Following nuclease treatment, the target:oligonucelotide hybrid complexes which are perfectly complementary are identified. From the location of the labelled targets, the oligonucleotide probes which hybridized with the targets can be identified and, in turn, the sequence of the target nucleic acid can be determined.

The particular nuclease used will depend on the target nucleic acid being sequenced. If the target is RNA, a RNA nuclease is used. Similarly, if the target is DNA, a DNA nuclease is used. RNase A is an example of an RNA nuclease that can be used to increase the quality of RNA hybridization signals on high density oligonucleotide arrays. RNase A effectively recognizes and cuts single-stranded RNA, including RNA in RNA:DNA hybrids that is not in a perfect double-stranded structure. Moreover, RNA bulges, loops, and even single base mismatches can be recognized and cleaved by RNase

A. In addition, RNase A recognizes and cleaves target RNA which binds to multiple oligonucleotide probes present on the substrate. S1 nuclease and Mung Bean nuclease are examples of DNA nucleases which can be used to improve the DNA hybridization signals on high density oligonucleotide arrays. Other nucleases, which will be apparent to those of skill in the art, can similarly be used to increase the quality of RNA hybridization signals on high density oligonucleotide arrays and, in turn, to more accurately determine the sequence of the target nucleic acid.

FIG. 4 is a schematic outline of a hybridization procedure which can be carried out prior to nuclease treatment. Fluorescein-UTP and -CTP labelled RNA is prepared from a PCR product by *in vitro* transcription. The RNA is fragmented by heating and allowed to hybridize with an array of oligonucleotide probes on a single substrate. The array of oligonucleotide probes is generated using the tiling procedure described so that the array of oligonucleotide probes is capable of recognizing substantially all of the possible subsequences present in the target RNA. Moreover, for purposes of comparison, the array of oligonucleotides is preferably generated so that all of the four possible probes for a given position to be identified are in close proximity to one another (*i.e.*, so that they are in predefined regions which are near to one another). Following hybridization, the substrate is rinsed with the hybridization buffer and a quantitative fluorescence image of the hybridization pattern is obtained by, for example, scanning the substrate with a confocal microscope. It should be noted that confocal detection allows hybridization to be measured in the presence of excess labelled target and, hence, if desired, hybridization can be detected in real time.

Following hybridization, the substrate having an array of target: oligonucleotide hybridization complexes thereon is contacted with a nuclease. Frequently, this is carried out by flowing a solution of the nuclease over the substrate using, for example, techniques similar to the flow channel methods described above. The nuclease solution is typically formed using the buffer used to carry out the hybridizaiton reaction (*i.e.*, the hybridization buffer). The concentration of the nuclease will vary depending on the particular nuclease used, but will typically range from about 0.05 $\mu$g/ml to about 2 mg/ml. Moreover, the time in which the array of target:oligonucleotide hybridization complexes is in contact with the nuclease will vary. Typically, nuclease treatment is carried out for a period of time ranging from about 5 minutes to 3 hours. Following treatment with the nuclease, the substrate is again washed

with the hybridization buffer, and a quantitative fluorescence image of the hybridization pattern is obtained by scanning the substrate with, for example, a confocal microscope.

As such, nuclease treatment can be used following hybridization to improve the quality of hybridization signals on high density oligonucleotide arrays and, in turn, to more accurately determine the sequence of the target nucleic acid. It will be readily apparent to the those of skill in the art that the foregoing is inteded to illustrate, and not restrict, the way in which an array of target:oligonucelotide hybrid complexes can be treated with a nuclease to improve hybridization signals on high density oligonucleotide arrays.

### *B. Enhanced Discrimination Using Ligation Reactions*

Ligiation reactions can be used to discriminate between fully complementary hybrids and those that differ by one or more base pairs. More particularly, an array of oligonucelotides is generated on a substrate (in the 3' to 5' direction) using any one of the methods described above. Each of the oligonucleotides in the array is shorter in length than the target nucleic acid so that when hybridized to the target nucleic acid, the target nucleic acid generally has a 3' overhang. In this embodiment, the target nucleic acid is not necessarily labelled. After the array of oligonucelotides has been combined with the target nucelic acid to form target-oligonucleotide hydrid complexes, the target-oligonucleotide hybrid complexes are contacted with a ligase and a labelled, ligatable probe or, alternatively, with a pool of labelled, ligatable probes. The ligation reaction of the labelled, ligatable probes to the 5' end of the oligonucleotide probes on the substrate will occur, in the presence of the ligase, only when the target:oligonucelotide hybrid has formed with correct base-pairing near the 5' end of the oligonucleotide probe and where there is a suitable 3' overhang of the target nucleic acid to serve as a template for hybridization and ligation. After the ligation reaction, the substrate is washed (multiple times if necessary) with water at a temperature of about 40°C to 50°C to remove the target nucleic acid and the labelled, unligated probes. Thereafter, a quantitative fluorescence image of the hybridization pattern is obtained by scanning the substrate with, for example, a confocal microscope, and labelled oligonucleotide probes, *i.e.*, the oligonucleotide probes which are perfectly complementary to the target nucleic acid, are identified. Using this information, the sequence of the target nucleic acid can be determined.

Any enzyme that catalyzes the formation of a phosphodiester bond at the site of a single-strand break in duplex DNA can be used to enhace discrimination between fully complementary hybrids and those that differ by one or more base pairs. Such ligases include, but are not limited to, T4 DNA ligase, ligases isolated from *E. coli* and ligases isolated from other bacteriophages. The concentration of the ligase will vary depending on the particular ligase used, the concentration of target and buffer conditions, but will typically range from about 500 units/ml to about 5,000 units/ml. Moreover, the time in which the array of target:oligonucleotide hybridization complexes is in contact with the ligase will vary. Typically, the ligase treatment is carried out for a period of time ranging from about 2 hours to 200 hours.

As such, litigation reactions can be used to improve discrimination of base-pair mismatches near the 5' end of the immobilized probe, mismatches that are often poorly discriminated following hybridization alone. It will be readily apparent to those of skill in the art that the foregoing is inteded to illustrate, and not restrict, the way in which an array of target:oligonucelotide hybrid complexes can be treated with a ligase and a pool of labelled, ligatable probes to improve hybridization signals on high density oligonucleotide arrays.

## *VI. General Hybridization Parameters*

The extent of specific interaction between oligonucleotide probes immobilized to the VLSIPS substrate and another sequence specific reagent may be modified by the conditions of the interaction. Sequencing embodiments typically require high fidelity hybridization and the ability to discriminate perfect matching from imperfect matching. Fingerprinting and mapping embodiments may be performed using less stringent conditions, or in some embodiments very highly stringent conditions, depending upon the circumstances.

In a nucleic acid hybridization embodiment, the specificity and kinetics of hybridization have been described in detail by, *e.g.*, Wetmur ana Davidson *J. Mol. Biol.*, 31:349-370 (1968), Britten and Kohne *Science* 161:529-530 (1968), and Kanehisa, *Nuc. Acids Res.* 12:203-213 (1984), each of which is hereby incorporated herein by reference. Parameters which are well known to affect specificity and kinetics of reaction include salt conditions, ionic composition of the solvent, hybridization temperature, length of oligonucleotide matching sequences, guanine and cytosine (GC) content,

presence of hybridization accelerators, pH, specific bases found in the matching sequences, solvent conditions, and addition of organic solvents.

Generally, the salt concentration will depend on the stringency desired. The typical salt used is sodium chloride (NaCl), however, other ionic salts may be utilized, *e.g.*, KCl. Depending on the desired stringency of hybridization, the salt concentration will often be less than about 3 molar, more often less than 2.5 molar, usually less than about 2 molar, and more usually less than about 1.5 molar. For applications directed towards higher stringency matching, the salt concentrations would typically be lower. Ordinary high stringency conditions will utilize salt concentration of less than about 1 molar, more often less then about 750 millimolar, usually less than about 500 millimolar, and may be as low as about 250 or 150 millimolar.

The kinetics of hybridization and the stringency of hybridization both depend upon the temperature at which the hybridization is performed and the temperature at which the washing steps are performed. Temperatures at which steps for low stringency hybridization are desired would typically be lower temperatures, *e.g.*, ordinarily at least about 5°C, more ordinarily at least about 20°C, usually at least about 25°C, and more usually at least about 30°C. For those applications requiring high stringency hybridization, or fidelity of hybridization and sequence matching, temperatures at which hybridization and washing steps are performed would typically be high. For example, temperatures in excess of about 35°C and up to 45°C may often be used. Of course, the hybridization of oligonucleotides may be disrupted by even higher temperatures. Thus, for stripping of targets from substrates, as discussed below, temperatures as high as 80°C, or even higher may be used.

The base composition of the specific oligonucleotides involved in hybridization affects the temperature of melting, and the stability of hybridization as discussed in the above references. However, the bias of GC rich sequences to hybridize faster and retain stability at higher temperatures can be compensated for by the inclusion in the hybridization incubation or wash steps of various buffers. Sample buffers which accomplish this result include CTAB and the triethly-and trimethyl ammonium buffers. *See, e.g.*, Wood, *et al.* (1987) Proc. Natl. Acad. Sci. USA, 82:1585-1588, and Khrapko, K., *et al.* (1989) FEBS Letters 256:118-122.

The rate of hybridization can also be affected by the inclusion of particular hybridization accelerators. These hybridization accelerators include the volume exclusion

agents characterized by dextran sulfate, or polyethylene glycol (PEG). Dextran sulfate is typically included at a concentration of between 1% and 40% by weight. The actual concentration selected depends upon the application, but typically a faster hybridization is desired in which the concentration is optimized for the system in question. Dextran sulfate is often included at a concentration of between 0.5% and 2% by weight or dextran sulfate at a concentration between about 0.5% and 5%. Alternatively, proteins which accelerate hybridization may be added, *e.g.*, the recA protein found in *E. coli*) or other homologous proteins.

Of course, the specific hybridization conditions will be selected to correspond to a discriminatory condition which provides a positive signal where desired but where the signal is considerably lower for probes not matched to target. This may be determined by a number of titration steps or with a number of controls which will be run during the hybridization and/or washing steps to determine at what point the hybridization conditions have reached the stage of desired specificity.

## *VI. Detection Methods*

Methods for detection depend upon the label selected. The criteria for selecting an appropriate label are discussed below, however, a fluorescent label is preferred because of its extreme sensitivity and simplicity. Standard labeling procedures are used to determine the positions where interactions between a target sequence and a reagent take place. For example, if a target sequence is labeled and exposed to a matrix of different oligonucleotide probes, only those locations where the oligonucleotides interact with the target will exhibit any signal. In addition to using a label, other methods may be used to scan the matrix to determine where interaction takes place. The spectrum of interactions can, of course, be determined in a temporal manner by repeated scans of interactions which occur at each of a multiplicity of conditions. However, instead of testing each individual interaction separately, a multiplicity of sequence interactions may be simultaneously determined on a matrix.

### *A. Labeling Techniques*

The target nucleic acid can be labeled using any of a number of convenient detectable markers. A fluorescent label is preferred because it provides a very strong signal with low background. It is also optically detectable at high resolution and

sensitivity through a quick scanning procedure. Other potential labeling moieties include, radioisotope, chemiluminescent compounds, labeled binding proteins, heavy metal atoms, spectroscopic markers, magnetic labels, and linked enzymes.

In another embodiment, different targets can be simultaneously sequenced where each target has a different label. For instance, one target could have a green fluorescent label and a second target could have a red fluorescent label. The scanning step will distinguish cites of binding of the red label from those binding the green fluorescent label. Each sequence can be analyzed independently from one another.

Suitable chromogens which can be employed include those molecules and compounds which adsorb light in a distinctive range of wavelengths so that a color can be observed or, alternatively, which emit light when irradiated with radiation of a particular wave length or wave length range, *e.g.*, fluorescers.

A wide variety of suitable dyes are available, being primary chosen to provide an intense color with minimal absorption by their surroundings. Illustrative dye types include quinoline dyes, triarylmethane dyes, acridine dyes, alizarine dyes, phthaleins, insect dyes, azo dyes, anthraquinoid dyes, cyanine dyes, phenazathionium dyes, and phenazoxonium dyes.

A wide variety of fluorescers can be employed either by alone or, alternatively, in conjunction with quencher molecules. Fluorescers of interest fall into a variety of categories having certain primary functionalities. These primary functionalities include 1- and 2-aminonaphthalene, p,p'-diaminostilbenes, pyrenes, quaternary phenanthridine salts, 9-aminoacridines, p,p'-diaminobenzophenone imines, anthracenes, oxacarbocyanine, marocyanine, 3-aminoequilenin, perylene, bisbenzoxazole, bis-p-oxazolyl benzene, 1,2-benzophenazin, retinol, bis-3-aminopyridinium salts, hellebrigenin, tetracycline, sterophenol, benzimidzaolylphenylamine, 2-oxo-3-chromen, indole, xanthen, 7-hydroxycoumarin, phenoxazine, salicylatc, otrophanthidin, porphyrins, triarylmethanes and flavin. Individual fluorescent compounds which have functionalities for linking or which can be modified to incorporate such functionalities include, *e.g.*, dansyl chloride; fluoresceins such as 3,6-dihydroxy-9-phenylxanthhydrol; rhodamineisothiocyanate; N-phenyl 1-amino-8-sulfonatonaphthalene; N-phenyl 2-amino-6-sulfonatonaphthalene: 4-acetamido-4-iscthiocyanato-stilbene-2,2'-disulfonic acid; pyrene-3-sulfonic acid; 2-toluidinonaphthalene-6-sulfonate; N-phenyl, N-methyl 2-aminoaphthalene-6-sulfonate; ethidium bromide; stebrine;

auromine-0,2-(9'-anthroyl)palmitate; dansyl phosphatidylethanolamine; N,N'-dioctadecyl oxacarbocyanine; N,N'-dihexyl oxacarbocyanine; merocyanine, 4(3'pyrenyl)butyrate; d-3-aminodesoxy-equilenin; 12-(9'anthroyl)stearate; 2-methylanthracene; 9-vinylanthracene; 2,2'(vinylene-p-phenylene)bisbenzoxazole; p-bis[2-(4-methyl-5-phenyl-oxazolyl)]benzene; 6-dimethylamino-1,2-benzophenazin; retinol; bis(3'-aminopyridinium) 1,10-decandiyl diiodide; sulfonaphthylhydrazone of hellibrienin; chlorotetracycline; N(7-dimethylamino-4-methyl-2-oxo-3-chromenyl)maleimide; N-[p-(2-benzimidazolyl)-phenyl]maleimide; N-(4-fluoranthyl)maleimide; bis(homovanillic acid); resazarin; 4-chloro-7-nitro-2,1,3benzooxadiazole; merocyanine 540; resorufin; rose bengal; and 2,4-diphenyl-3(2H)-furanone.

Desirably, fluorescers should absorb light above about 300 nm, preferably about 350 nm, and more preferably above about 400 nm, usually emitting at wavelengths greater than about 10 nm higher than the wavelength of the light absorbed. It should be noted that the absorption and emission characteristics of the bound dye can differ from the unbound dye. Therefore, when referring to the various wavelength ranges and characteristics of the dyes, it is intended to indicate the dyes as employed and not the dye which is unconjugated and characterized in an arbitrary solvent.

Fluorescers are generally preferred because by irradiating a fluorescer with light, one can obtain a plurality of emissions. Thus, a single label can provide for a plurality of measurable events.

Detectable signal can also be provided by chemiluminescent and bioluminescent sources. Chemiluminescent sources include a compound which becomes electronically excited by a chemical reaction and can then emit light which serves as the detectible signal or donates energy to a fluorescent acceptor. A diverse number of families of compounds have been found to provide chemiluminescence under a variety or conditions. One family of compounds is 2,3-dihydro-1,-4-phthalazinedione. The must popular compound is luminol, which is the 5-amino compound. Other members of the family include the 5-amino-6,7,8-trimethoxy- and the dimethylamino[ca]benz analog. These compounds can be made to luminesce with alkaline hydrogen peroxide or calcium hypochlorite and base. Another family of compounds is the 2,4,5-triphenylimidazoles, with lophine as the common name for the parent product. Chemiluminescent analogs include para-dimethylamino and -methoxy substituents. Chemiluminescence can also be obtained with oxalates, usually oxalyl active esters, *e.g.*, p-nitrophenyl and a peroxide,

*e.g.*, hydrogen peroxide, under basic conditions. Alternatively, luciferins can be used in conjunction with luciferase or lucigenins to provide bioluminescence.

Spin labels are provided by reporter molecules with an unpaired electron spin which can be detected by electron spin resonance (ESR) spectroscopy. Exemplary spin labels include organic free radicals, transitional metal complexes, particularly vanadium, copper, iron, and manganese, and the like. Exemplary spin labels include nitroxide free radicals.

### *B. Scanning System*

With the automated detection apparatus, the correlation of specific positional labeling is converted to the presence on the target of sequences for which the oligonucelotides have specificity of interaction. Thus, the positional information is directly converted to a database indicating what sequence interactions have occurred. For example, in a nucleic acid hybridization application, the sequences which have interacted between the substrate matrix and the target molecule can be directly listed from the positional information. The detection system used is described in PCT publication no. WO90/15070; and U.S.S.N. 07/624,120. Although the detection described therein is a fluorescence detector, the detector can be replaced by a spectroscopic or other detector. The scanning system can make use of a moving detector relative to a fixed substrate, a fixed detector with a moving substrate, or a combination. Alternatively, mirrors or other apparatus can be used to transfer the signal directly to the detector. *See, e.g.*, U.S.S.N. 07/624,120, which is hereby incorporated herein by reference.

The detection method will typically also incorporate some signal processing to determine whether the signal at a particular matrix position is a true positive or may be a spurious signal. For example, a signal from a region which has actual positive signal may tend to spread over and provide a positive signal in an adjacent region which actually should not have one. This may occur, *e.g.*, where the scanning system is not properly discriminating with sufficiently high resolution in its pixel density to separate the two regions. Thus, the signal over the spatial region may be evaluated pixel by pixel to determine the locations and the actual extent of positive signal. A true positive signal should, in theory, show a uniform signal at each pixel location. Thus, processing by plotting number of pixels with actual signal intensity should have a clearly

uniform signal intensity. Regions where the signal intensities show a fairly wide dispersion, may be particularly suspect and the scanning system may be programmed to more carefully scan those positions.

More sophisticated signal processing techniques can be applied to the initial determination of whether a positive signal exists or not. *See, e.g.*, U.S.S.N. 07/624,120.

From a listing of those sequences which interact, data analysis may be performed on a series of sequences, for example, in a nucleic acid sequence application, each of the sequences may be analyzed for their overlap regions and the original target sequence may be reconstructed from the collection of specific subsequences obtained therein. Other sorts of analyses for different applications may also be performed, and because the scanning system directly interfaces with a computer the information need not be transferred manually. This provides for the ability to handle large amounts of data with very little human intervention. This, of course, provides significant advantages over manual manipulations. Increased throughput and reproducibility is thereby provided by the automation of vast majority of steps in any of these applications.

## *VII. DATA ANALYSIS*

Data analysis will typically involve aligning the proper sequences with their overlaps to determine the target sequence. Although the target "sequence" may not specifically correspond to any specific molecule, especially where the target sequence is broken and fragmented up in the sequencing process, the sequence corresponds to a contiguous sequence of the subfragments.

The data analysis can be performed manually or, preferably, by a computer using an appropriate program. Although the specific manipulations necessary to reassemble the target sequence from fragments may take many forms, one embodiment uses a sorting program to sort all of the subsequences using a defined hierarchy. The hierarchy need not necessarily correspond to any physical hierarchy, but provides a means to determine, in order, which subfragments have actually been found in the target sequence. In this manner, overlaps can be checked and found directly rather than having to search throughout the entire set after each selection process. For example, where the oligonucleotide probes are 10-mers, the first 9 positions can be sorted. A particular subsequence can be selected as in the examples, to determine where the process starts.

As analogous to the theoretical example provided above, the sorting procedure provides the ability to immediately find the position of the subsequence which contains the first 9 positions and can compare whether there exists more than 1 subsequence during the first 9 positions. In fact, the computer can easily generate all of the possible target sequences which contain given combination of subsequences. Typically, there will be only one, but in various situations, there will be more.

Generally, such computer programs provides for automated scanning of the substrate to determine the positions of oligonucleotide and target interaction. Simple processing of the intensity of the signal may be incorporated to filter out clearly spurious signals. The positions with positive interaction are correlated with the sequence specificity of specific matrix positions, to generate the set of matching subsequences. This information is further correlated with other target sequence information, *e.g.*, restriction fragment analysis. The sequences are then aligned using overlap data, thereby leading to possible corresponding target sequences which will, optimally, correspond to a single target sequence.

## *VII. Applications*

The technology provided by the present invention has very broad applications. Although described specifically for polynucleotide sequences, similar sequencing, fingerprinting, mapping, and screening procedures may be applied to polypeptide, carbohydrate, or other polymers. This may be for de novo sequencing, or may be used in conjunction with a second sequencing procedure to provide independent verification. See, *e.g.*, *Science* 242:1245 (1988). For example, a large polynucleotide sequence defined by either the Maxam and Gilbert technique or by the Sanger technique may be verified by using the present invention.

In addition, by selection of appropriate probes, a polynucleotide sequence can be fingerprinted. Fingerprinting is a less detailed sequence analysis which usually involves the characterization or a sequence by a combination of defined features. Sequence fingerprinting is particularly useful because the repertoire of possible features which can be tested is virtually infinite. Moreover, the stringency of matching is also variable depending upon the application. A Southern Blot analysis may be characterized as a means of simple fingerprint analysis.

Fingerprinting analysis may be performed to the resolution of specific nucleotides, or may be used to determine homologies, most commonly for large segments. In particular, an array of oligonucleotide probes of virtually any workable size may be positionally localized on a matrix and used to probe a sequence for either absolute complementary matching, or homology to the desired level of stringency using selected hybridization conditions.

In addition, the present invention provides means for mapping analysis of a target sequence or sequences. Mapping will usually involve the sequential ordering or a plurality of various sequences, or may involve the localization of a particular sequence within a plurality of sequences. This may be achieved by immobilizing particular large segments onto the matrix and probing with a shorter sequence to determine which of the large sequences contain that smaller sequence. Alternatively, relatively shorter probes of known or random sequence may be immobilized to the matrix and a map of various different target sequences may be determined from overlaps. Principles of such an approach are described in some detail by Evans et al. (1989) "Physical Mapping of Complex Genomes by Cosmid Multiplex Analysis," *Proc. Natl. Acad. Sci. USA* 86:5030-5034; Michiels, *et al.*, "Molecular Approaches to Genome Analysis: A Strategy for the Construction of Ordered Overlap Clone Libraries," *CABIOS* 3:203-210 (1987); Olsen, *et al.* "Random-Clone Strategy for Genomic Restriction Mapping in Yeast," *Proc. Natl. Acad. Sci. USA* 83:7826-7830 (1986); Craig, *et al.*, "Ordering of Cosmid Clones Covering the Herpes Simplex Virus Type I (HSV-I) Genome: A Test Case for Fingerprinting by Hybridization," *Nuc. Acids Res.* 18:2653-2660 (1990); and Coulson, *et al.*, "Toward a Physical Map of the Genome of the Nematode Caenorhabditis elegans," *Proc. Natl. Acad. Sci. USA* 83:7821-7825 (1986); each of which is hereby incorporated herein by reference.

Fingerprinting analysis also provides a means of identification. In addition to its value in apprehension of criminals from whom a biological sample, *e.g.*, blood, has been collected, fingerprinting can ensure personal identification for other reasons. For example, it may be useful for identification of bodies in tragedies such as fire, flood, and vehicle crashes. In other cases the identification may be useful in identification of persons suffering from amnesia, or of missing persons. Other forensics applications include establishing the identity of a person, *e.g.*, military identification "dog tags", or may be used in identifying the source of particular biological samples. Fingerprinting

technology is described, *e.g.*, in Carrano, *et al.*, "A High-Resolution, Fluorescence-Based, Semi-automated method for DNA Fingerprinting," *Genomics* 4: 120-136 (1989), which is hereby incorporated herein by reference.

The fingerprinting analysis may be used to perform various types of genetic screening. For example, a single substrate may be generated with a plurality of screening probes, allowing for the simultaneous genetic screening for a large number of genetic markers. Thus, prenatal or diagnostic screening can be simplified, economized, and made more generally accessible.

In addition to the sequencing, fingerprinting, and mapping applications, the present invention also provide, means for determining specificity of interaction with particular sequences. Many of these applications are described in U.S.S.N. 07/362,901 (VLSIPS parent), U.S.S.N. 07/492,462 (VLSIPS CIP), U.S.S.N. 07/435,316 (caged biotin parent), and U.S.S.N. 07/612,671 (caged biotin CIP), which are incorporated herein by reference.

## *X. Examples*

The following examples are provided to illustrate the efficacy of the inventions herein.

### *A. ENHANCED DISCRIMINATION USING RNase A*

This example illustrates the ability of RNase A to recognize and cut single-stranded RNA, including RNA in DNA:RNA hybrids that is not in a perfect double-stranded structure. RNA bulges, loops, and even single base mismatches can, for example, be recognized and cleaved by RNase A. RNase A treatment is used herein to improve the quality of RNA hybridization signals on high density oligonucleotide arrays.

#### ***EXAMPLE I***

The high density array of oligonucleotide probes on a glass substrate (referred to as a "chip") is prepared using the standard VLSIPS protocols set forth above. Moreover, the pattern of oligonucleotide probes is based on the standard tiling strategy described shown in Fig. 5. Briefly, the chip used in this example consists of an overlapping set of DNA 15-mers covalently linked to a glass surface. A set of four probes for each nucleotide of a 1.3 kb region spanning the D-loop region of human mitochondrial DNA (mtDNA) is present on the chip. Each of the four probes contains a different base (A, C, G or T) at the position being interrogated, with the substitution position being near the center of the probe. Because the probes are specifically selected based on the mtDNA target sequence, one of the four probes will be perfectly complementary to the mtDNA target, and the other three will contain a central base-pairing mismatch. The mismatch probes are expected to hybridize to a lesser extent. By incorporating a fluorophore into the target DNA or RNA, the extent of hybridization at the four positions for each base can be quantitated using fluorescence imaging. In principle, the correct target base is simply identified as the complement to the probe base giving rise to the largest hybridization signal.

Generally, a "base identification" is considered to be made if the signal in one of the four probe regions is greater than twice as large as the signal in a nearby region that contains no oligonucleotide probes (referred to herein as the "background"), ***and*** if the signal is at least 1.2 times as large as in the other three related probe regions on the chip. If the signal in more than one of the probe regions is larger than twice the

background, but is not greater than the other three by at least a factor of 1.2, then a "multiple-base ambiguity" is indicated. For example, if the T-containing and the C-containing probes have high but similar hybridization signals, a two-base ambiguity would result (a call of either the complementary bases A or G could be made). All two-base ambiguities are possible, as well as all 3- and 4-base ambiguities. If the most intense hybridization signal (largest by at least a factor of 1.2) is in the region that is not complementary to the target sequence, then an "incorrect call" is made (referred to herein as a "miscall"). As shown below, the RNase A treatment resolves multiple-base ambiguities and reduces the number of miscalls that result from hybridization of a 1.3 kb RNA target to the mitochondrial probe chip described above.

Labelled mitochondrial RNA samples are prepared using standard PCR and *in vitro* transcription procedures. The 1.3 kb RNA sample is labelled by incorporation of fluorescein-labelled UTP during transcription (approximately 10% of Us in the RNA sample are labelled). The RNA (approximately 200 nM concentration of 1.3 kb transcripts) is partially fragmented by heating to 99.9°C for 60 minutes in 6 mM magnesium chloride, pH 8. This procedure produces a wide range of fragment lengths, with an average length of approximately 200 nucleotides. After fragmentation, the RNA sample is diluted to 10 nM in 60 mM sodium phosphate, 0.9 M NaCl, 6 mM EDTA, 0.05% Triton X-100, pH 7.9 (referred to as 6XSSPE-T). For hybridization, 10 mM CTAB (cetyltrimethylammonium bromide) is added. The RNA sample is hybridized to the chip in a 1 ml flow cell at 22°C for 40 minutes with stirring provided by bubbling nitrogen gas through the flow cell. Following hybridization, the chip is rinsed with 6XSSPE-T and the fluorescence signal is detected using a scanning confocal fluorescence microscope ("reading" the chip). (*See*, FIG. 6) The image is stored for later analysis. The chip is then treated with 75 $\mu$l of 0.2 $\mu$g/ml RNase A in 6XSSPE-T at 22°C for intervals of 10, 45, and 75 minutes. After each interval, the chip is rinsed with 6XSSPE-T and the fluorescence signal is read. (*See*, FIG. 7) The results are analyzed to determine the number of correct base calls, multiple-base ambiguities and miscalls, and the improvement resulting from the RNase A treatments.

After the original hybridization, 619 out of 1302 bases were called correctly (approximately 47%). Of the remaining, there were 218 miscalls, 458 multiple-base ambiguities, and 17 instances where the signal was not more than twice the background. (These numbers are subject to the conditions of the experiment.) In

particular, they are a function of hybridization time and temperature, salt concentration, the presence of Triton X-100 and CTAB, and the extent of RNA fragmentation and labelling. The conditions used here, in particular the limited fragmentation of the RNA, are ones that tend to decrease the number of regions with low signal, and to increase the number of miscalls and ambiguities.) Following treatment with RNase A (and combining the information for the three time points), 162 out of 218 miscalls were corrected, and 350 out of 458 ambiguities were correctly resolved. There were only 46 bases that were initially ambiguous which were resolved incorrectly, and there were no instances of correct calls that were changed to incorrect calls after RNase A treatment. After the initial hybridization, only 47%, of the entire sequence was called correctly. However, when the hybridization results are combined with the results following RNase A treatment, approximately 87% of the 1302 bases are called correctly. These results clearly demonstrate that RNase A is very effective in improving the quality of the sequence information obtained from hybridization to oligonucleotide arrays.

### *B. ENHANCED DISCRIMINATION USING LIGATION REACTIONS*

The following examples illustrate the ability of ligation reactions to improve discrimination of base-pair mismatches near the 5' end of an oligonucleotide probe. The ligation reaction of labelled, short oligonucleotides to the 5' end of oligonucleotide probes on a chip should occur (in the presence of the enzyme Ligase) wherever a probe:target hybrid has formed with correct base-pairing near the 5' end of the probe and where there is a suitable 3' overhang of the target to serve as a template for hybridization and ligation. In the following examples, the ligation reaction is used to improve discrimination of base-pair mismatches near the 5' end of the probe, *i.e.*, mismatches which are often poorly discriminated following hybridization alone.

***Example I***

In this example, a chip is made with probes having the following sequence:

P-P-A-A-CGCGCCGCNC-5' (Seq. ID No.:2)

wherein: P is a polyethlyeneglycol (PEG) spacer, A, C, and G, are the usual deoxynucleotides, and N is either A, C, G, or T. The chip is made using the standard VLSIPS protocols set forth above. The target oligonucleotide is a 20-mer having the following sequence (listed 5' to 3'):

F1-GCGCGGCGCGAACGCAACGC (Seq. ID No.:3)

wherein: F1 is a fluorescein molecule covalently attached at the 5' end. The labelled, ligatable 6-mer used in this example has the following sequence:

F1-TGCGTT.

The 5' half of the 20-mer target is complementary to the probes on the chip for which N is a G. The probe:target hybrids for the other three probes have a single base mismatch one base in from the 5' end of the probe. The ligatable 6-mer is complementary to the 3' overhang of the target when the target is hybridized to the probe to form the maximum number of Watson-Crick hydrogen bonds.

Prior to hybridization and ligation, the chip is treated with T4 Polynucleotide Kinase in order to phosphorylate the 5' end of the probes. The probes are phosphorylated using 100 units of T4 Polynucleotide Kinase (New England Biolabs) in 1 ml at 37°C for 90 minutes.

A 10 nM solution of the target oligo in 6XSSP-T (no EDTA in the hybridization buffer because EDTA could interfere with subsequent ligation reactions) is hybridized to the chip for 30 minutes at 22°C. The chip is scanned, and then washed with a large amount of water to remove the labelled target molecules.

The ligation reaction is carried out at 16°C in a 1 ml flow cell containing 10 nM target oligo, 20 nM ligatable 6-mer, and 4000 units of T4 DNA Ligase (New England Biolabs). The buffer is the buffer recommended by the manufacturer plus 150 mM NaCl. The reaction is allowed to proceed for 14 hours at 16°C, after which the

chip is vigorously washed with water at 50°C to remove the labelled target molecules. The only fluorescent label remaining after washing is that of the ligatable 6-mers that have been covalently attached to the probes via the ligation reaction. The chip is scanned and analyzed, and the results compared to those obtained from the hybridization reaction above.

| N | HYB | HDF | LIG | LDF |
|---|---|---|---|---|
| A | 143 | 1.1 | 15 | 5.5 |
| C | 134 | 1.1 | 13 | 6.3 |
| G** | 151 | 1.0 | 82 | 1.0 |
| T | 110 | 1.4 | 20 | 4.1 |

In the above table, N is the base in the probe that is one position in from the 5' end (*see, supra*). For the target used here, G is the complementary base. HYB and LIG are the signals (fluorescence counts) for the different probes following hybridization and ligation, respectively. HDF and LDF are the discrimination factors (defined as the ratio of the fluorescence signal with the perfect match, G, to the signal with the specified mismatch base) following hybridization and ligation, respectively.

It is clear that after hybridization, the extent of target hybridization is very similar for the perfectly complementary probe and the probes containing a mismatch near the 5' end. The A and C mismatches differ by only 10%, and the maximum difference is only 40%. In contrast, following the ligation reaction, the discrimination is greatly improved, with the minimum discrimination factor greater than 4. These data indicate that ligation reactions can be performed on covalently attached oligonucleotide probes on the chip surface, that these reactions are specific for correctly base-paired probe:target hybrids, and that the reaction can be used to improve the discrimination between perfect matches and single base mismatches.

## EXAMPLE II

In this example, a chip was made with probes having the following sequences:

P-P-A-A-CGCGCATTCN-5' (denoted CG) Seq. ID No.:4
P-P-A-A-ATATAATTCN-5' (denoted AT) Seq. ID No.:5

A, T, C, G and N have the same definitions as those set forth in Example I, *supra*. These probes contain a perfect match and the single-base mismatch sequences for the following 22-mer target oligos (listed 5' to 3'):

F1-GCGCGTAAGGCCTTCGACGTAG (denoted OH1) Seq. ID No:6
F1-TATATTAAGGCCTTCGACGTAG (denoted OH2) Seq. ID No:7

The 5' end of OH1 is complementary to the CG probes with N = C, and the 5' end of OH2 is complementary to the AT probes with N = C. Both OH1 and OH2 have the same 12-mer sequence at the 3' end. The labelled, ligatable 6-mer used in this example (appropriate for both OH1 and OH2 when hybridized to the CG and AT regions of the chip, respectively) has the following sequence:

F1-CGAAGG (denoted L6B).

Prior to hybridization and ligation, the chip is phosphorylated as in Example I, *supra*, using T4 polynucleotide kinase for 4 hours at 37°C. The hybridization and ligation conditions are the same as those used in Example I unless otherwise specified. In particular, 2000 units of T4 DNA Ligase are used for the reaction here, and the concentration of the ligatable 6-mer is 10 nM rather than 20 nM.

The hybrids between OH1 and the CG probes on the chip contain a high proportion of C-G base pairs. C-G base pairs are known to be considerably more stable than the A-T base pairs that are predominant in the hybrid between OH2 and the AT probes on the chip. Thus, it is expected that OH1 will hybridize to its perfectly complimentary probe oligo to a greater extent than will OH2 under suitably stringent hybridization conditions. In fact, this is observed to be the case in the hybridization experiments below. The ligation reaction, however, can be used to help mitigate the complicating effects of the base composition dependence of hybridization.

The chip was initially hybridized with both OH1 and OH2 at 22°C for 30 minutes. The extent of hybridization to both the CG and AT regions of the chip is analyzed. It is found that the fluorescence signal in the CG regions (OH1 hybrids) is larger than in the AT regions (OH2 hybrids) by more than a factor of 14. In fact, the perfect match signal in the CG region is quite strong, but the signal in the AT region is only slightly greater than twice the background.

| | (OH1) | | (OH2) | |
|---|---|---|---|---|
| N | HYB | HDF | HYB* | HDF* |
| A | 196 | 2.4 | 6 | 5.5 |
| C** | 474 | 1.0 | 33 | 1.0 |
| G | 159 | 3.0 | 20 | 1.7 |
| T | 103 | 4.6 | 5 | 6.6 |

* These values are somewhat uncertain because the signal is not large relative to the background.

Following hybridization, the chip was washed extensively with water to remove the target molecules. A ligation reaction is initiated on the chip by combining OH1, OH2, and L6B in 1 ml of ligation buffer and adding 2000 units of T4 DNA Ligase. The reaction is allowed to proceed for 34 hours at 22°C, and then for another 24 hours at 8°C. At each stage, the chip is read and the data recorded and analyzed.

| | 34 hrs., T = 22°C | | | | 24 hrs., T = 8°C | | | |
|---|---|---|---|---|---|---|---|---|
| | (OH1]) | | (OH2) | | (OH1) | | (OH2) | |
| N | LIG | LDF | LIG | LDF | LIG | LDF | LIG | LDF |
| A | 18 | 56 | 3 | 31 | 27 | 46 | 10 | 88 |
| C | 1003 | 1.0 | 92 | 1.0 | 1234 | 1.0 | 879 | 1.0 |
| G | 13 | 44 | 23 | 13 | 24 | 51 | 30 | 29 |
| T | 15 | 67 | 3 | 31 | 22 | 56 | 8 | 110 |

It is striking that after the ligation reaction at 8°C, the signals for OH1 and OH2 differ by only a factor of 1.4, ten times less than the factor of 14 that was observed following the original hybridization. It is even more striking that the composition dependence is mitigated by virtue of the ligation reaction at low temperature with no loss of discrimination for either OH1 or OH2.

***Example III***

In order for the ligation strategy to be useful for unknown or more complex DNA targets, it is necessary to use a pool of all possible (4096) 6-mers instead of a specific ligatable 6-mer. The 4096 6-mers are synthesized using standard phosphoramidite

chemical procedures on four separate columns, one beginning (at the 3' end) with A, one with C, one with G, and one with T. Each of the 5 subsequent synthesis steps are performed using a mixture of A, C, G, and T phosphoramidite, producing a mixture of all possible five base sequences on each of the four columns. The 6-mers are labelled with fluorescein at the 5' end as the last step in the synthesis. After reversed-phase HPLC purification of the four 6-mer pools, the concentration of each pool is determined by the absorption at 260 nm. The appropriate amounts of each pool is mixed to make a solution that contains all 4096 labelled 6-mer oligonucleotides.

A chip is made containing 10-mer probes having the following sequences

$$P\text{-}P\text{-}C\text{-}G\text{-}C\text{-}G\text{-}N_1\text{-}N_2\text{-}N_3\text{-}N_4\text{-}N_5\text{-}N_6\text{-}5'$$ (Seq. ID No. 8)

wherein: $N_i$ are A, C, G, or T. In other words, the chip contains 10-mers with all possible (4096) six base combinations at the 5' end. The 5' phosphate group on the probes required for ligation is added chemically (using 5' Phosphate-ON, Clontech Laboratories, Palo Alto, CA) as the last step in the synthesis of the chip, prior to deprotection of the bases. The target oligo is a 22-mer having the following sequence (listed 5' to 3'):

F1-GCGCGTAAGGCCTTCGACGTAG (OH1)

The chip was initially hybridized with 10 nM OH1 in 6XSSP-T at 22°C for 30 minutes. The chip is read and analyzed. The only perfect match probe for this target (*i.e.*, PP-CGCGCATTCC-5') (Seq. ID No. 9) has the second highest hybridization signal. Eight other probes have hybridization signals that are within a factor of 4 of the perfect match signal. The other three probes with a single base mismatch at the 5' end have discrimination factors of 2.0, 2.6, and 3.5, for G, A, and T, respectively. Other single base mismatches at positions in from the 5' end of the probe give signals that are considerably smaller. The chip is washed with water to remove the hybridized target.

The chip is next hybridized using the conditions used for the ligation reaction. The chip is hybridized with 10 nM OH1 and 1.6 $\mu$M 6-mer pool (0.4 nM for each 6-mer oligo) in the ligation buffer for 11 hours at 22°C (no ligase at this stage). The perfect match probe gives the highest signal by a factor of 2.4. Five probes have signals within a factor of 4 of the perfect match signal. The other three probes with a single base mismatch at the 5' end have discrimination factors of 3.0, 3,6, and 8.0, for G, A, and T, respectively.

The ligation reaction is initiated by the addition of 2000 units of T4 DNA ligase to the solution containing OH1 and the pool of 6-mers. The reaction is allowed to proceed for 23 hours at 22°C. After washing the chip with water at about 45°C for five minutes, the chip is read. After ligation, no other probes have hybridization signals that are within a factor of 4 of the perfect match signal. The three 5' single base mismatch probes all have discrimination factors greater than 12. Thus, with a complex chip containing 4096 probes with all possible 6-mer sequences at the 5' end, and using a pool of all possible ligatable 6-mers, the ligation reaction is still specific for the perfectly complementary probe and affords considerable increases in the discrimination between perfect matches and single-base mismatches.

## EXAMPLE IV

In this example, a chip was made using the tiling strategy (A, C, G, T -containing probes for each base in the sequence) described above that covers a 50 base region of the protease gene of HIV-1 (SF2 strain). The probes are 11-mers, linked to the glass support by three PEG linkers. The substitution position (the position being interrogated by an A, C, G, or T base in the probe) is varied between the 5' end of the probe, and five bases in from the 5' end (referred to as positions end, -1, -2, -3, -4 and -5). The chip is synthesized using standard VLSIPS protocols. Prior to hybridization and ligation, the chip is phosphorylated using T4 polynucleotide kinase for 5 hours at 37°C. The target is a 75-mer oligonucleotide (denoted Hpro1), labelled at the 5' end with fluorescein, that spans the complementary 50 base region on the chip.

The chip was initially hybridized with a 10 nM solution of Hpro1 in 6XSSP-T at 22°C for 30 minutes. After hybridization, the chip was read, and then rinsed with water to remove the target molecules. A ligation reaction was then carried out with 10 nM Hpro1, 1.6 $\mu$M 6-mer pool (0.4 nM per oligo), and 2000 units of T4 DNA Ligase in 1 ml of ligation buffer. The ligation reaction is allowed to proceed for 25 hours at 8°C, then 90 hours at 22°C, and finally 4 days at 8°C. At intervals of 1 to 2 days, the solution is supplemented with additional T4 DNA Ligase. Following the ligation reaction, the chip is washed vigorously with water at about 45°C for 10 minutes, leaving only the labelled 6-mers that have been ligated to the probe molecules. The chip is read, and the data analyzed.

The results of the hybridization and ligation reactions are analyzed in terms of the ability to make a correct base call from the fluorescence signal measured on the chip. In particular, the signal is compared between the four probes that differ by a single base at a given position within the 11-mer, with the rest of the 11-mer being perfectly complementary to a specific region of the target sequence. For the purposes of this experiment, a base identification is said to be made if the signal in at least one of the four probe regions is greater than the signal in a nearby region that has no oligonucleotide probes (the background) by at least 5 counts (the background counts are usually about 2 - 6 counts), ***and*** if the signal in one of the four regions is greater than that in the other three related regions by at least a factor of 1.2. If none of the four signals are larger than the other three by a factor of at least 1.2, a multiple base ambiguity results. If the most intense hybridization signal (by a factor of at least 1.2) is for a probe that is not perfectly complementary to the target sequence, then a miscall results.

Following hybridization, the 11-mer probes with substitution positions -1, -2, -3, and -4 all gave 49 correct base calls and 1 multiple base ambiguity. The probe with substitution position -5 resulted in 50 correct base calls. Following ligation, the probes with substitution positions -2 and -5 gave 48 correct calls and 2 miscalls, substitution position -3 yielded 48 correct calls and 1 ambiguity and 1 miscall, and substitution position -1 and -4 both yielded 50 correct calls with no ambiguities or miscalls. These results indicate that the ligation reaction with the full pool of 6-mers can be used to specifically label hybrids between relatively complex targets and arrays of oligonucleotide probes.

It is interesting to note that the pattern of ligation (stronger or weaker signals, better or worse discrimination) is not in general the same as the pattern of hybridization. This suggests that these two approaches may be used as complementary tools to obtain sequence information with arrays of oligonucleotide probes. For example, probes that produce large hybridization signals, but are poorly discriminated may be better treated using a ligation step. And probes that do not hybridize well to a particular complementary target (leading to a signal that is too small relative to the background) may ligate well enough to be clearly detected (as also suggested by the mitigation of the base composition dependence demonstrated in Example II, *supra*).

## *XI. Conclusion*

The present invention provides greatly improved methods and apparatus for the study of nucleotide sequences and nucleic acid interactions with other molecules. It is to be understood that the above description is intended to be illustrative and not restrictive. Many embodiments and variations of the invention will become apparent to those of skill in the art upon review of this disclosure. Merely by way of example, certain of the embodiments described herein will be applicable to other polymers, such as peptides and proteins, and can utilized other synthesis techniques. The scope of the invention should, therefore, be determined not with reference to the above description, but instead should be determined with reference to the appended claims along with the full scope of equivalents to which such claims are entitled.